Introduction

Brazil is the biggest producer of sugarcane in the world. There are different climates, relief and vegetation in Brazil which results to diversity of soil types and differences in levels of soil fertility. To improve and sustain sugarcane production capacity of Brazil, it is pertinent to assess fertility of soils from regions that are major players in sugarcane cultivation by analyzing soil physicochemical properties. The results of the physicochemical analyses will indicate the type of soil fertility treatment suitable for each sugarcane cultivation zone. This project used multivariate statistical techniques to group soil samples from different regions of Brazil into clusters based on similarity of physicochemical properties. This grouping makes it easier to design specific soil fertility restoration treatment for each cluster.

Description of the dataset

Since almost all the sugarcane crops in Brazil are produced in the South-Central and NorthEast regions (Nassar et al., 2018), the soil samples in the data set used for this report were obtained from 27 locations in 9 states located in South-Central and NorthEast regions of Brazil. The soils were collected from 0 to 20 cm layer. The number of samples collected from any given state varied according to the heterogeneity of soils in that region (Chagas et al., 2019). The dataset contains 27 observations and 17 variables (of which three variables are non-numeric). In the dataset, the name of the numeric variables were written as chemical symbols and the full meaning are as follows:

1. V - Base saturation
2. CTC - Effective cation exchange capacity
3. Ca^{2+} - Calcium ions
4. M - Saturation by aluminum
5. MO - Organic carbon
6. Al^{3+}- Aluminum ion
7. P – Phosphorus
8. K^{+} – Potassium ion
9. Mg^{2+} - Magnesium ion
10. pH – measure of acidity or alkalinity
11. H+Al – Exchangeable acidity
12. Sandy – sand content of soil
13. Silt – Silt content of soil
14. Clay – clay content of soil.

Choice of methodology and motivation

The objective of this project is to group the soil samples into clusters based on similarity of physicochemical properties. Since there are 14 variables representing the physicochemical properties of the soil samples, there is need to reduce the number of variables and use only important variables in the clustering analysis. Based on Pearson test which showed that there is correlation among the variables, principal component analysis was selected as a dimension reduction technique to generate new uncorrelated variables that maximally explain variance in the data. The scores of these principal components were used in other multivariate techniques employed later in this project. Also, principal component analysis is a good choice for dimension reduction because the variables are numeric and continuous.

To achieve the aim of this project, non-hierarchical clustering method (k-means) was used to form clusters of soil samples. Euclidean distance between the physicochemical properties of the soil samples was used as similarity measure. Clustering analysis is an unsupervised technique and it is suitable for grouping the soil samples based on similarity as we do not know beforehand the appropriate number of clusters for the soil samples nor patterns in the data. The optimal number of seeds used in k-means clustering was obtained using minimization of the within-clusters sum of squares criterion. Finally, linear discriminant analysis was used to confirm and validate clustering analysis. Linear discriminant analysis is a suitable technique for this purpose since it can classify observations into multiple categories based on their properties which validates the accuracy of clustering analysis. Also, LDA produces a predictive model for classifying new observations into known groups or groups generated via clustering analysis.

R-code

```
library(factoextra); library (GGally); library(tidyverse); library (MASS); library(MVN)
### read data
soil_data = read.csv ("soilData.csv", header = TRUE) # set working directory to folder with data set
soil = soil_data [ ,4:17]                                   # select the numeric columns or variables
rownames(soil) = soil_data [, 3]                  # set row names to Soil IDs
plot(soil)                        # graphically explore data set and check for outliers / correlation
soil = soil [-9,]                                           # remove outlier
### Principal component analysis
ggcorr(soil, label = TRUE)            # graphical Pearson correlation result
pca = princomp(soil, cor = TRUE)      # principal component function
screeplot(pca, type = "lines")        # pick 4 principal components
pca$loadings                          # the contribution of variables to each principal component
summary(pca)                          # the first four PCs explains ~ 80% of the variance
pca$sdev                              # standard deviation of PCs
pca.var = pca$sdev^2                  # eigenvalues of the principal components
pca$scores                            # values of the principal components / new variables
fviz_pca_biplot (pca)                 # biplot of PCA
fviz_pca_var (pca)                    # graph of variables
PCs = as.data.frame(pca$scores[,1:4]) # a new data frame of the scores of the  principal components
### k means clustering
my_data = scale(PCs [,1:4])                         # scales data to eliminate effect of variance
fviz_nbclust(my_data, kmeans, method = "wss")       # compute total within-cluster sum of square
groups = kmeans(PCs, centers = 4, nstart = 25)      # group into clusters
fviz_cluster(groups, data = PCs)                    # cluster plot
### Linear discriminant analysis
mvn(my_data,mvnTest = "mardia")                     # test of multivariate normality
PCs$Clusters = as.factor(groups$cluster)            # cluster results to PCs data frame
soil.lda=lda(Clusters~.,PCs)                  # fits an LDA model
soil.lda                                      # shows details of LDA model
plot(soil.lda, col = as.integer(PCs$Clusters))  # plot of LDA model
soil.pred=predict(soil.lda)                   # predict cluster of observations using LDA model
soil.pred                                     # shows posterior probability of observations
hist(soil.pred$x)                             # histogram of projections
table(PCs$Clusters,soil.pred$class,dnn=c("From","Classified into"))     # accuracy table
soil.ldacv = lda(Clusters~.,PCs, CV=TRUE)    # evaluation of LDA model with cross-validation
table(PCs$Clusters,soil.ldacv$class,dnn=c("From","Classified into")) # accuracy of LDA
```

R-output

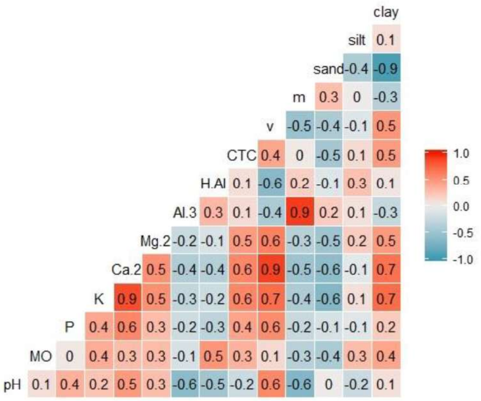


Figure 1: Pairwise correlation plot

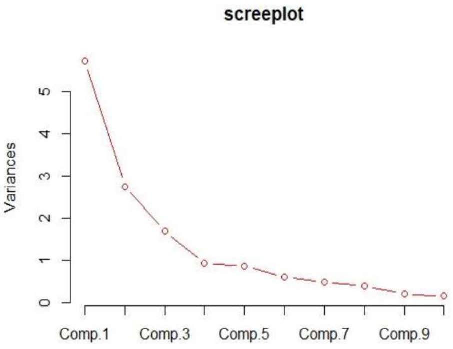
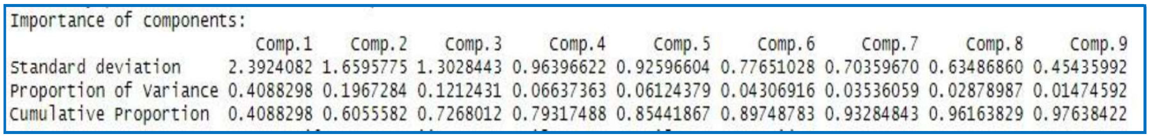
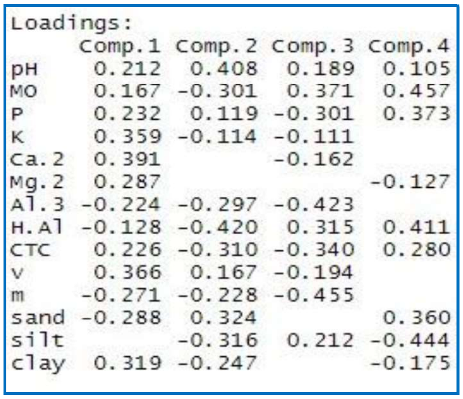


Figure 2: Scree plot of PCA



```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8     Comp.9
Standard deviation     2.3924082 1.6595775 1.3028443 0.96396622 0.92596604 0.77651028 0.70359670 0.63486860 0.45435992
Proportion of Variance 0.4088298 0.1967284 0.1212431 0.06637363 0.06124379 0.04306916 0.03536059 0.02878987 0.01474592
Cumulative Proportion  0.4088298 0.6055582 0.7268012 0.79317488 0.85441867 0.89748783 0.93284843 0.96163829 0.97638422
```

Table 1: Summary report of PCA



```
Loadings:
     Comp.1 Comp.2 Comp.3 Comp.4
pH    0.212  0.408  0.189  0.105
MO    0.167 -0.301  0.371  0.457
P     0.232  0.119 -0.301  0.373
K     0.359 -0.114 -0.111       
Ca.2  0.391        -0.162       
Mg.2  0.287               -0.127
Al.3 -0.224 -0.297 -0.423       
H.Al -0.128 -0.420  0.315  0.411
CTC   0.226 -0.310 -0.340  0.280
v     0.366  0.167 -0.194       
m    -0.271 -0.228 -0.455       
sand -0.288  0.324         0.360
silt        -0.316  0.212 -0.444
clay  0.319 -0.247        -0.175
```

Table 2: Loadings of principal components

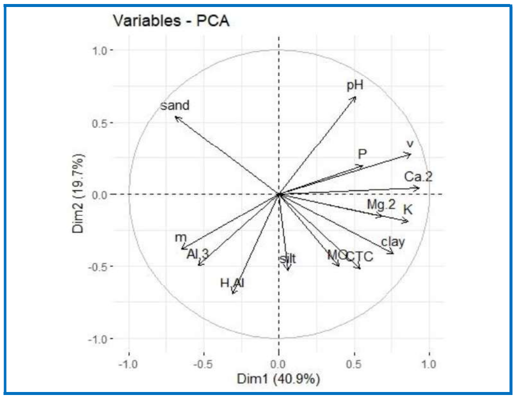


Figure 3: Biplot showing variables

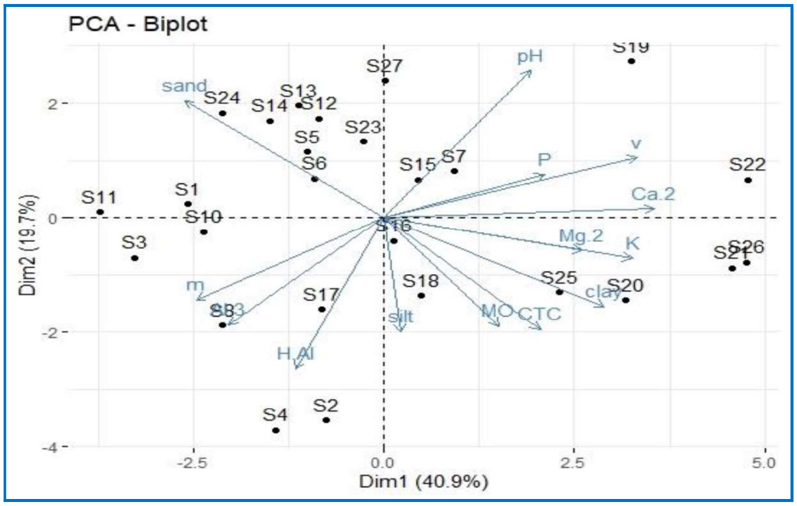
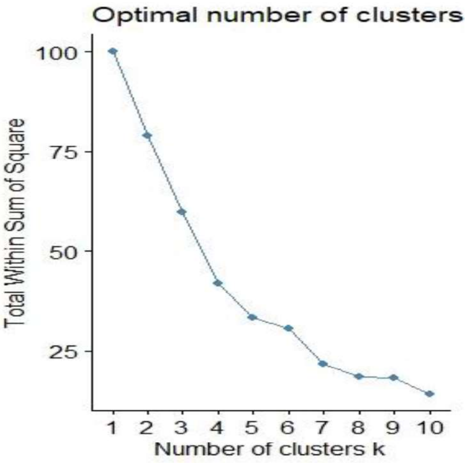


Figure 4: Biplot showing variables and observations



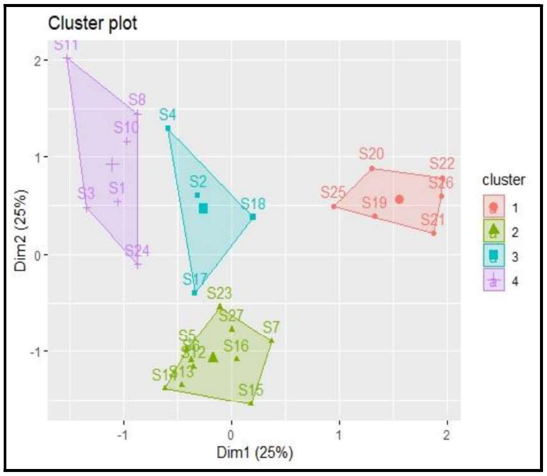


Figure 5: WSS plot for optimal number of clusters

Figure 6: K – means clustering

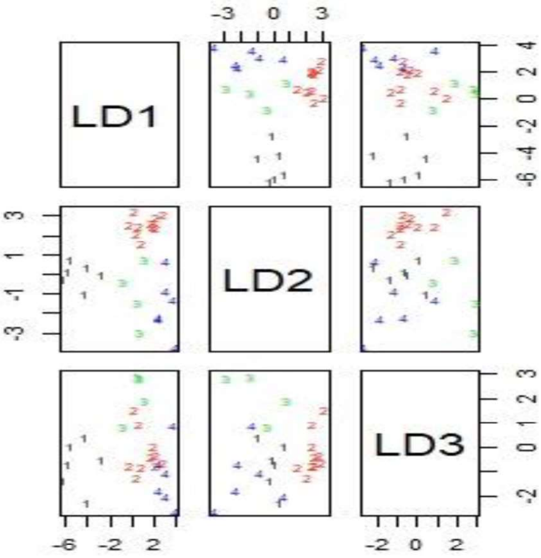
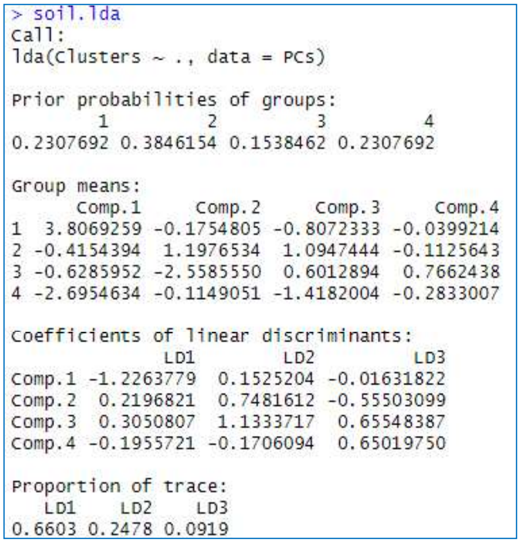
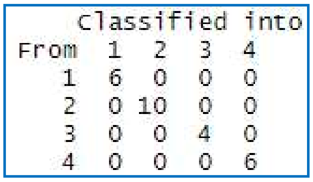


Figure 7: plot of LDA model



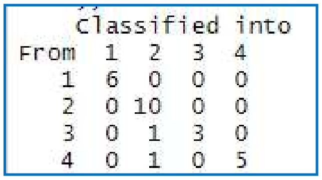
```
> soil.lda
Call:
lda(Clusters ~ ., data = PCs)

Prior probabilities of groups:
        1         2         3         4 
0.2307692 0.3846154 0.1538462 0.2307692 

Group means:
      Comp.1     Comp.2     Comp.3     Comp.4
1  3.8069259 -0.1754805 -0.8072333 -0.0399214
2 -0.4154394  1.1976534  1.0947444 -0.1125643
3 -0.6285952 -2.5585550  0.6012894  0.7662438
4 -2.6954634 -0.1149051 -1.4182004 -0.2833007

Coefficients of linear discriminants:
              LD1        LD2         LD3
Comp.1 -1.2263779  0.1525204 -0.01631822
Comp.2  0.2196821  0.7481612 -0.55503099
Comp.3  0.3050807  1.1333717  0.65548387
Comp.4 -0.1955721 -0.1706094  0.65019750

Proportion of trace:
   LD1    LD2    LD3 
0.6603 0.2478 0.0919 
```

Figure 8: summary of LDA model



```
      Classified into
From  1  2  3  4
   1  6  0  0  0
   2  0 10  0  0
   3  0  0  4  0
   4  0  0  0  6
```

Figure 9: LDA classification accuracy (100%)



```
      Classified into
From  1  2  3  4
   1  6  0  0  0
   2  0 10  0  0
   3  0  1  3  0
   4  0  1  0  5
```

Figure10: Accuracy of LDA with cross-validation *(92%)*

Interpretation of Results

Principal component analysis (PCA): As shown in Figure 1, there is correlation between the variables which necessitates the use of PCA as a dimension reduction technique to select variables that maximally and independently explain the variance in the sample. Based on the scree plot result (Figure 2), the first four principal components were selected for further use in the project. These four principal components cumulatively explain approximately 80% of the variance in the dataset (Table 1). The PCA loadings (Table 2) result revealed variables that had the highest contribution to each of the four principal components. These variables are Ca^{2+}, pH, MO and P for principal components 1, 2, 3 and 4 respectively. PCA biplot (figure 3 and figure 4) shows the spatial relationship between the observations and variables.

Clustering analysis: The result of K-means clustering (figure 6) exposed four clusters that broadly correspond to soil textures which are dependent on the content of clay, silt and sand in the soil. Also, the clusters were distinctive in respect to the soil nutrients found in each of the soil textures. Based on the PCA biplot, the clusters have these characteristics:

ID	Clusters	Texture	Chemical properties
1	S2, S4, S17 and S18	Silt	Organic carbon, and H+Al
2	S19, S20, S21, S22, S25 and S26	Clay	MO, CTC, P, K, V, Mg^{2+} and Ca^{2+}
3	S1, S3, S8, S10, S11 and S24	Silty sand	Saturated with Aluminium
4	S5, S6, S7, S12, S13, S14, S15, S16, S23 and S27	Sand	High pH

Linear discriminant analysis (LDA): The clustering result was added as a categorical variable to the data frame of principal component scores and it was used as the response variable in LDA. The LDA model had 100% predictive accuracy (figure 9) whereas cross validation of the LDA model had 92% predictive accuracy with two misclassification errors (figure 10). Since, 92% accuracy of LDA model is within acceptable range, it implies that the clustering analysis is satisfactory too.

The prior probabilities of groups (figure 8) indicate the probability of a random observation belonging to any of the four clusters. Group means represent the average value of predictor variables in each cluster. The coefficients are the parameters used to fit the predictive linear discriminant functions for classification of new observations into clusters. The linear discriminant functions are as follows:

Linear discriminant 1: cluster = $-1.2264(Ca^{2+}) + 0.2197(pH) + 0.3051(MO) – 0.1956(P)$

Linear discriminant 2: cluster = $0.1525(Ca^{2+}) + 0.7482(pH) + 1.1334(MO) – 0.1706(P)$

Linear discriminant 3: cluster = $-0.0163(Ca^{2+}) - 0.5550(pH) + 0.6555(MO) + 0.6502(P)$

The proportion of trace indicates the between-class variance that is explained by successive discriminant functions. LD1 explains 66% while LD2 explains 25% of between-class variance respectively. The summary of the predict function displays the posterior probability which is the probability of classifying each observation into a cluster based on available data.

Discussion and conclusion

It can be inferred that geographical locations in cluster 2 have the most fertile soil because they contain essential nutrients and have high CTC (cation exchange capacity) which indicates the volume of nutrients that can be held by the soil. The capability of clay soils to hold moisture and retain plant nutrients contributes to the high fertility of soil samples in this cluster. Likewise, presence of calcium ions in the clay soils will increase flocculation which increases soil fertility. Cluster 1 is composed of silty soils which are relatively fertile since they have some level of organic carbon content. Cluster 3 is composed of silty sand soils with high levels of aluminium which reduces soil fertility. Cluster 4 is composed of sandy soils with no nutrients because sandy soils drain well without retaining water and nutrients.

Based on the physicochemical attributes of each of the clusters, soil fertility remediation programs can be designed to improve sustainable sugar cane production. For example, the fertility of soil samples in cluster 4 can be improved by addition of Biochar. Likewise, the fertility of soil samples in cluster 3 can be enhanced by increasing the pH and addition of organic matter to reduce aluminium content.

In conclusion, the results obtained in this project will be a useful guide in improving the soil fertility of sugar cane plantations in Brazil.

References

Chagas, P.S.F.d., Souza, M.d.F., Dombroski, J.L.D. et al. Multivariate analysis reveals significant diuron-related changes in the soil composition of different Brazilian regions. Sci Rep 9, 7900 (2019) doi:10.1038/s41598-019-44405-x

André Meloni Nassar, Bernardo F.T. Rudor, Laura Barcellos Antoniazzi, Daniel Alves de Aguiar, Miriam Rumenos Piedade Bacchi and Marcos Adami. http://sugarcane.org/wp-content/uploads/2018/04/Wageningen-Chapter-3.pdf.